Prediction of Organic Reaction Outcomes Using Machine Learning

ACS central science

Research Article

http://pubs.acs.org/journal/acscii

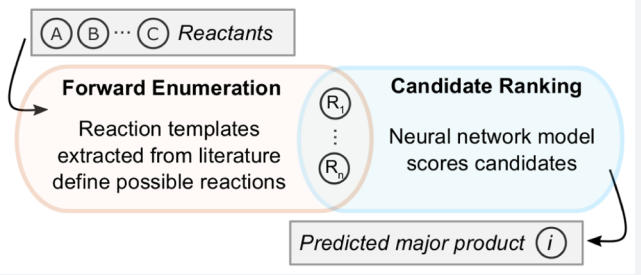
Prediction of Organic Reaction Outcomes Using Machine Learning

Connor W. Coley,[†] Regina Barzilay,[‡] Tommi S. Jaakkola,[‡] William H. Green,*[,†] and Klavs F. Jensen*[,†]

[†]Department of Chemical Engineering and [‡]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

Supporting Information

ABSTRACT: Computer assistance in synthesis design has existed for over 40 years, yet retrosynthesis planning software has struggled to achieve widespread adoption. One critical challenge in developing high-quality pathway suggestions is that proposed reaction steps often fail when attempted in the laboratory, despite initially seeming viable. The true measure of success for any synthesis program is whether the predicted outcome matches what is observed experimentally. We report a model framework for anticipating reaction outcomes that combines the traditional use of reaction templates with the flexibility in pattern recognition afforded by neural networks. Using 15 000 experimental reaction records from granted United States patents, a model is trained to select the major (recorded) product by ranking a self-generated list of candidates where one candidate is known to be the major product. Candidate reactions are represented using a unique edit-based representation that emphasizes the fundamental transformation from reactants to products, rather than the constituent molecules' overall structures. In a 5-fold cross-validation, the trained model assigns the major product rank 1 in 71.8% of cases, rank ≤3 in 86.7% of cases, and rank ≤5 in 90.8% of cases.



INTRODUCTION

Synthesis planning is often referred to as an art. The process of identifying a suitable pathway (i.e., series of reaction steps) which transforms some set of available reactants into a target compound is typically performed by expert chemists with years or decades of experience. To assist chemists with this task, computer-aided synthesis design was introduced over 40 years ago in the form of retrosynthetic planning software.

Retrosynthesis was originally formalized by Corey and Wipke[1,2] in their efforts to introduce computer assistance to synthesis with Logic and Heuristics Applied to Synthetic Analysis (LHASA).[3] Corey's approach to codifying retrosynthesis involved the explicit identification of molecular structures which lend themselves to disconnection or, rather, can be produced by known reactions in the forward direction. Almost all approaches to automated retrosynthesis, LHASA included, involve the use of reaction templates—submolecular patterns that encode changes in atom connectivity. Recursively applying retrosynthetic templates to a target molecule produces a candidate synthesis tree. However, a synthetic route based on retrosynthetic templates does not always lead to a successful forward synthesis. Templates are locally defined pattern-matching rules, inherently naive to what is present in the rest of the molecule. It is common, therefore, for a proposed retrosynthetic disconnection to be unviable in the forward direction. Once a synthetic route has been proposed, it is critical to evaluate each step in the forward direction to identify these challenges.

Forward analysis is such an important part of pathway evaluation that even the very first retrosynthesis program, Corey's LHASA, could identify functional group conflicts which might lead to a lack of specificity or selectivity.[3] Another early program, Computer-Assisted Mechanistic Evaluation of Organic Reactions (CAMEO),[4] implemented a similar approach where nucleophilic and electrophilic sites were analyzed pairwise to determine qualitative reactivities. Other programs like SOPHIA[5] and Eros[6] identify potentially reactive functional groups using manually curated reactivity rules and empirical calculations. Chematica's Syntaurus[7] contains explicitly encoded lists of incompatible functional groups for each retrosynthetic template.

Manual encoding of these rules has obvious disadvantages. First, it relies on the intuition and experience of a small number of chemists. Second, it is not scalable—it is not realistic to exhaustively define the full substrate scope and incompatibilities for every possible reaction. Third, conflicting reactivity is rarely black and white; incompatibility depends on the exact nature of the reacting molecules. These factors motivate the development of an automated approach to forward reaction evaluation.

Work by Kayala et al. considers the problem of forward synthesis mechanistically, rather than using end-to-end templates.[8,9] They use graph-based representations of molecules and assign approximate molecular orbitals to each so that a mechanistic step can be considered an interaction between a donor and an acceptor orbital. Although they show very promising results, the need for manual encoding of mechanistic rules to generate training data could be problematic and limit

Received: February 3, 2017
Published: April 18, 2017

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443

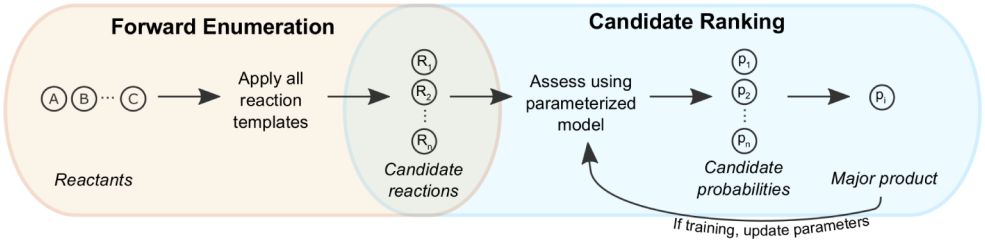


Figure 1. Model framework combining forward enumeration and candidate ranking. The primary aim of this work is the creation of the parametrized scoring model, which is trained to maximize the probability assigned to the recorded experimental outcome.

scalability. Wei et al.[10] describe the use of neural networks to predict the outcome of reactions based on reactant fingerprints, but limit their study to 16 types of reactions covering a very narrow scope of possible alkyl halide and alkene reactions. Given two reactants and one reagent, the model was trained to identify which of 16 templates was most applicable. The data set used for cross-validation comes from artificially generated examples with limited chemical functionality, rather than experimental data.

Quite recently, Segler and Waller describe two approaches to forward synthesis prediction. The first is a knowledge-graph approach that uses the concept of half reactions to generate possible products given exactly two reactants by looking at the known reactions in which each of those reactants participates.[11] In one validation, the recorded product is found in the list of candidate products (with a median size of 3, mean 5.3) 67.5% of the time using a knowledge-base of eight million reactions. With "only" one million reactions, performance drops to just 15%. The second approach uses neural networks to rank reaction templates given reactant fingerprints. For reactions that are known to be contained in their automatically extracted set of 8720 templates, the authors report accuracies as high as 78%, but do not quantify the coverage of their template set; an average of just 44.5 matches per query suggests poor coverage.[12] The data used for training and testing are not precisely defined, nor is the code/model available for comparative purposes.

Previous studies have not relied on published experimental reaction examples due to the challenge of only having "positive" examples, excepting Segler and Waller.[12] The literature is heavily biased toward reactions with high yields, and reactions with negligible or zero yields are rarely reported except for illustrative purposes, e.g., highlighting the necessity of a catalyst. Proprietary electronic lab notebooks can contain numerous unproductive reactions, including results of high-throughput screening, but these data are not available in public or even commercial databases. This limitation of reaction databases precludes many supervised learning approaches for forward synthetic prediction: one cannot train a model directly on literature data to classify a certain reaction as productive or unproductive, since there are almost no unproductive examples available.

In this work, we describe a model that learns to predict the major products of chemical reactions given a set of reactant molecules by combining rigid reaction templates and machine learning. Specifically, we report the following contributions: (1) a data augmentation strategy whereby reaction databases are supplemented with chemically plausible negative reaction examples; (2) the successful application of that strategy using automatically extracted forward synthesis templates, where poor specificity is not a hindrance and no manual curation is required; (3) a new reaction representation focused on the fundamental transformation at the reaction site rather than constituent reactant and product fingerprints; (4) the implementation and validation of a neural network-based model that learns when certain modes of reactivity are more or less likely to occur than other potential modes. Despite the literature bias toward reporting only high-yielding reactions, we develop a successful workflow that can be performed without any manual curation using actual reactions reported in the USPTO literature.

APPROACH

Overview. Our model predicts the outcome of a chemical reaction in a two-step manner: (1) applying overgeneralized forward reaction templates to a pool of reactants to generate a set of chemically plausible products, and (2) estimating which candidate product is the major product as a multiway classification problem using machine learning. This is shown schematically in Figure 1.

In the first stage, we apply a library of forward synthetic templates to define which products could be produced based on the initial reactants. Rule-based enumeration has been applied to forward synthesis analysis previously,[13,14] but because many distinct templates can match a given reactant set and generate hundreds or thousands of products, simple enumeration is not inherently useful. If a particular reaction is plausible, but it proceeds at a rate insignificant compared to other reactions, then we do not need to consider it when evaluating the viability of a forward reaction step. In the second stage, each candidate reaction is scored individually by the machine learning model. This model assesses the likelihood of reactivity, akin to a reaction rate, in isolation from competing reactions. The scores from all candidates are compared in a softmax network layer (i.e., an exponential activation function that maps a list of numbers to a list of probabilities that sum to one) to generate probabilities describing which product is predicted to be most abundant.

A key component of our approach is this two-step formalization. By generating candidate products in the first step, in effect, existing reaction databases are augmented with negative reaction examples. This circumvents the limitation of only having high-yielding reaction data. Implicit in a reaction example A + B → C with greater than 50% yield is that A + B ↛ D, A + B ↛ E, etc., or at least that D and E were formed to a lesser extent than the reported C, where D and E are plausible alternate products. Including these alternate products in the training set allows us to extract more information from each reaction entry than we otherwise could by, e.g., training to predict the yield of only recorded reactions.

The recorded product of a reaction in the patent database is the "true" product that the model learns to predict, while the

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443

chemically plausible alternate products generated via templates are the "false" products which were not reported in the literature. A model can then be trained to identify the "true" product as a multiway classification or ranking problem. Treating recorded products as true outcomes is consistent with the literature bias toward reporting high-yield reactions.

Data. The source of reaction examples is the set of USPTO patents granted between 1976 and 2013, preparsed by Lowe.[15] Contextual information (e.g., temperature, solvent, reaction time) is inconsistently present, so reaction examples were reduced to reactants and products only. Forward templates were extracted from the prefiltered set of 1,122,662 atom-mapped reaction SMILES in 1976−2013_USPTOgrants_reactionSmiles_feb2014filters.rsmi.[16] For training and testing major product identification, a 15 000-member subset of the full set of reactions in 1976−2013_USPTOgrants_CML.7z[16] was used. These examples are formatted as CML documents where distinct roles (reactant, product, reagent, solvent, catalyst) are assigned to each molecule in the reaction. This data set was chosen over the previous one so that spectator molecules (e.g., reagents, solvents) could be given the opportunity to react, despite not contributing atoms to the reported products. No two reaction examples in the data set have an identical reactant pool.

Forward Enumeration. To build the database of forward templates, we use a heuristics-driven algorithm inspired by Law et al.[17] and Bogevig et al.[18] For each atom-mapped reaction example found in Lowe's USPTO database, the reaction core is defined by determining which product atoms have a different connectivity than the corresponding reactant atoms. The reaction core is expanded to include adjacent unmapped leaving groups and immediately neighboring atoms. Neighboring atoms are fully generalized into any non-hydrogen substituent for maximal generality to achieve high coverage at the expense of low specificity. A SMARTS string encoding the submolecular pattern at the reaction core can be generated for the reactants and for the products, which together define a reaction SMARTS string. A total of 140 284 unique reaction SMARTS strings are extracted from 1,122,662 reaction SMILES strings. Figure S1 shows the popularity of forward templates as a function of their rank. The five most popular templates are shown in Figure 2 with *ex post facto* labels. Although these five happen to be unimolecular functional group conversions, we have not predefined common functional groups or common transformations; moreover, the model does not rely on manual curation, labeling, or sorting of these extracted templates.

Due to imperfect canonicalization, the 140 284 templates contain some duplicates (e.g., same patterns with different numbering) and other redundancies (e.g., hydrolysis of ester overlaps with hydrolysis of alkyl ether/ester). Moreover, application of templates is computationally expensive (Figure S2), and the marginal coverage benefit of including additional templates decreases rapidly with rank (Figure S1). To focus the model on the most prevalent reaction types, only templates with more than 50 precedents were included in subsequent steps; this corresponds to the top 1689 templates.

For each reaction example, all molecules (reactants, reagents, catalysts, and solvents) were combined into a single reactant pool. The removal of annotations labeling species' roles was motivated by the fact that they were originally assigned using information about the recorded product. The 1689 templates were applied to the reactant pool to generate a list of potential

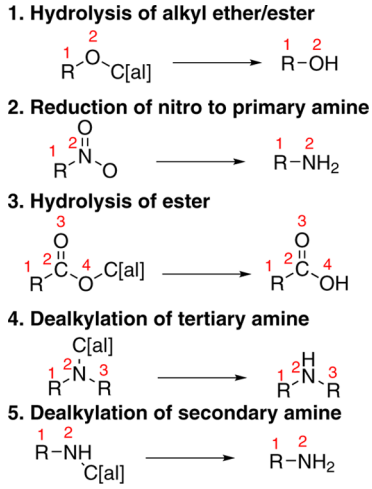


Figure 2. Depiction of the top five most popular forward synthetic templates extracted from 1.1 million USPTO reactions. C[al] denotes any aliphatic carbon.

products. The candidate products were each reduced to the product fragment with the longest SMILES string to neglect any byproduct salts and approximate the "major product". Atom mapping is preserved so that a candidate product corresponds to a fully atom-mapped candidate reaction. The inclusion of reagents, catalysts, and solvents makes the task of product prediction more challenging due to the competing candidate products they produce. The recorded product was found within the candidate sets in roughly 76% of reaction examples. The imperfect coverage (24%) reflects the use of the 1689 most popular templates rather than the full set of 140 284. The 15 000-member data set used for training and testing consists only of reactions where the recorded product was found within the candidate set generated by this template set. A histogram showing the number of candidate atom-mapped reactions for each example can be found in Figure S6. The peak occurs around 150 candidates with a median of 246 and mean of 353.

Prior to the subsequent candidate ranking step, during training, we determine which candidates match the "true" recorded reaction outcome. Atom-mapping is excluded from this comparison to limit the impact of its inaccuracies on model performance. When multiple candidates match the recorded product, the candidate corresponding to the most popular template is kept and the remaining matching candidates are discarded so only one "true" candidate exists.

Candidate Ranking. The parametrized reaction scoring function, a central component of the workflow in Figure 1, is challenging to design for many reasons, one of which is representation. A reaction is a complex data structure for which there is no universal vector-based description. We employ a new edit-based reaction representation strategy that emphasizes the change in atom connectivity that occurs at the reaction core during a chemical reaction. An atom-mapped reaction candidate is parsed into four different types of edits:

(1) An atom a_i loses a hydrogen
(2) An atom a_i gains a hydrogen
(3) Two atoms, a_i and a_j, lose a connecting bond b_{ij}
(4) Two atoms, a_i and a_j, gain a connecting bond b_{ij}

Changes in bond order are marked as a loss of the original bond order and a gain of the new bond order. With this

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443

representation, less fundamental changes (e.g., formal charge, association/disassociation of salts) are neglected. Loss or gain of a hydrogen is represented by 32 easy-to-compute features of that reactant atom alone, $a_i \in \mathbb{R}^{32}$. Loss or gain of a bond is represented by a concatenation of the features of the atoms involved and four features of the bond, $[a_i, b_{ij}, a_j] \in \mathbb{R}^{68}$.

Because edits occur at the reaction center by definition, the overall representation of a candidate reaction depends *only* on the atoms and bonds at the reaction core. There is no explicit inclusion of other molecular features, e.g., adjacency to certain functional groups. However, in our featurization, we include rapidly calculable structural and electronic features of the reactants' atoms that reflect the local chemical environment first and foremost, but also reflect the surrounding molecular context.[19−22] The chosen features can be found in Table 2 and Table S2 and are discussed in more detail in the Supporting Information.

The design of the neural network is motivated by the likelihood of a reaction being a function of the atom/bond changes that are required for it to occur. Individual edits are first analyzed in isolation using a neural network unique to the corresponding edit type. Three fully connected dense layers are used to embed initial features into an intermediate feature vector representation. The intermediate features vectors of all edits are summed and passed through a final neural network to produce a scalar score. This score, assigned to each candidate separately, represents the propensity of the proposed reaction (i.e., set of edits) to occur, akin to the negative free energy of reaction. The absolute likelihood scores of all candidates are compared in a final softmax layer, which produces a vector of probabilities from a vector of numbers by treating their values as pseudoenergies in a Boltzmann distribution with $k_B T = 1$. An example using simplified atom- and bond-level features is described in the Supporting Information for the reaction shown in Figure S4. When trained with this architecture, the four distinct neural networks become tailored to assessing their corresponding edit type. The architecture is depicted in Figure 3. Layer sizes were fixed prior to cross-validation tests after an initial screening to ensure sufficiently flexibility to describe the training data. All hidden layers are fully connected with bias and tanh activation. The final model required 144 001 parameters, described in Table S3.

We also implement a baseline model, which attempts to rank candidate products based on the products alone; no consideration is given to the reactants or corresponding reaction edits. For this model, product molecules are represented by radius-2 Morgan circular fingerprints of length 1024. A single hidden layer with tanh activation is used prior to the linear output layer. The baseline scoring model is shown in S5 with its 51 301 parameters (described in Table S4).

And finally, we implement a hybrid model, which trains the full edit-based and baseline architectures simultaneously and uses the sum of their scores for each candidate reaction.

RESULTS

Following the aforementioned procedures, 15 000 recorded reaction examples where the true product was found by applying the 1689 most popular templates were taken from the USPTO literature and augmented by adding the nonrecorded products to create a set of 5,335,669 examples. The model was trained and tested using a 5-fold cross-validation with the Adadelta optimizer[23] and early stopping. Each fold used a 70%/10%/20% training/validation/testing split and ceased training once the validation loss did not improve for five epochs. The edit-based model achieves an test accuracy of 68.5%, averaged across all folds. In this context, accuracy refers to the percentage of reaction examples where the recorded product was assigned a rank of 1. The baseline model was similarly trained and tested in a 5-fold CV, reaching an accuracy of 33.3%, suggesting that the set of recorded products in the data set is fairly homogeneous. The hybrid model, combining the edit-based representation with the proposed products' fingerprint representations, achieves an accuracy of 71.8%. These results are displayed in Table 1.

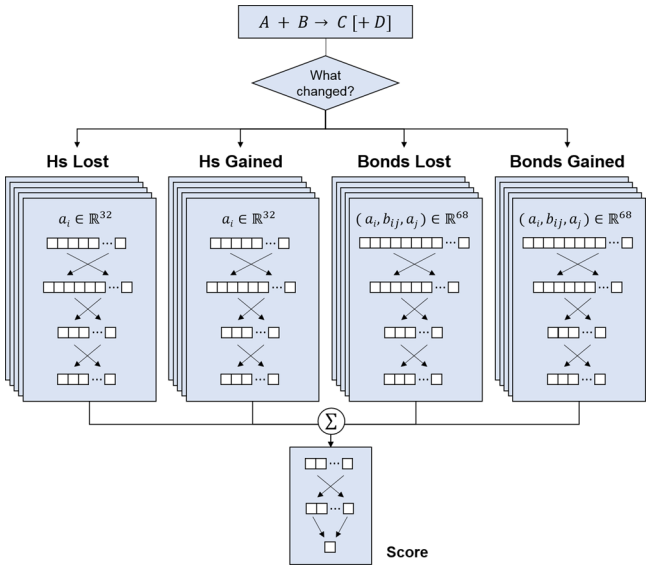


Figure 3. Edit-based model architecture for scoring candidate reactions. Reactions are represented by four types of edits. Initial atom- and bond-level attributes are converted into feature representations, which are summed and used to calculate that candidate reaction's likelihood score.

Table 1. Comparison between Baseline, Edit-Based, and Hybrid Models in Terms of Categorical Crossentropy Loss and Accuracy[a]

model	loss	acc. (%)	top-3 (%)	top-5 (%)	top-10 (%)
random guess	5.46	0.8	2.3	3.8	7.6
baseline	3.28	33.3	48.2	55.8	65.9
edit-based	1.34	68.5	84.8	89.4	93.6
hybrid	1.21	71.8	86.7	90.8	94.6

[a]Top-n refers to the percentage of examples where the recorded product was ranked within the top n candidates.

Accuracy is a simplified metric of model performance; the actual objective during training is minimization of the categorical crossentropy loss, $-\log p(x_{\text{true}})$, the average of the negative natural logarithm of the probability assigned to the true candidate. By this metric, which reflects both model accuracy and model confidence, the edit-based model (1.34) is vastly superior to the baseline model (3.28); the hybrid model (1.21) offers an additional improvement.

A direct comparison of prediction distributions is shown in Figure 4. As indicated by the baseline model's histogram of assigned probabilities in Figure 4a, the true outcome was assigned a near-zero probability in a majority of examples; in

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443

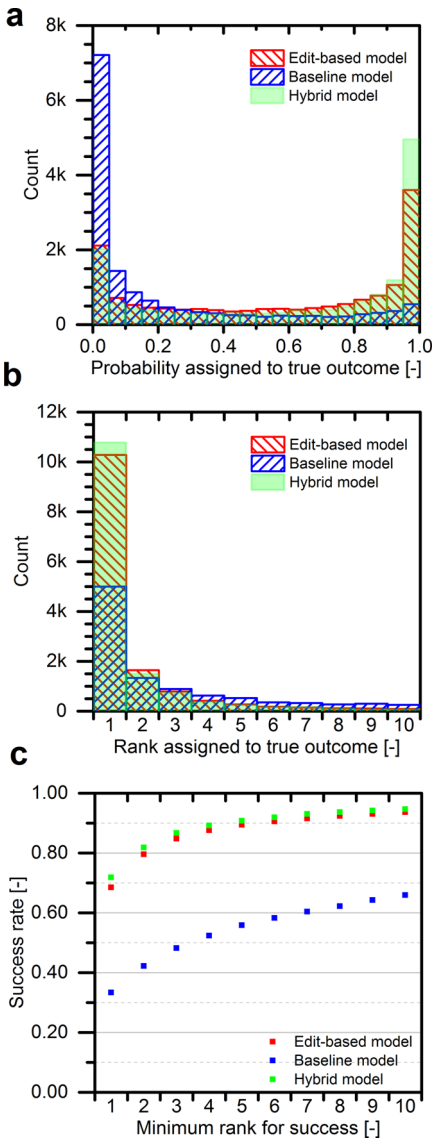


Figure 4. Performance of the three reaction prediction models as indicated by the (a) histogram of probabilities assigned to true outcomes; (b) histogram of ranks assigned to true outcomes, truncated to ranks 1−10; and (c) overall success rate as a function of the minimum acceptable assigned rank. In each case, the model is attempting to select the true product out of several hundred possible reaction products.

comparison, the edit-based model exhibits a more favorable distribution shifted toward higher probabilities—the hybrid model even more so. This is also reflected in Figure 4b, where the distribution of assigned ranks is short-tailed in the edit-based and hybrid models and long-tailed in the baseline model. The shape of this tail affects overall model success as the success criterion is relaxed from rank = 1 to rank ≥ *n*. Success rates of each model are shown in Figure 4c as a function of the minimum acceptable rank (i.e., the top-*n* accuracy). Figure S12 in the Supporting Information shows model performance as a function of how similar recorded products are to others in the data set.

It is important to understand the significance of the probability the model assigns to each candidate. Figure 5 depicts the performance of each model as a function of that model's "confidence". In this context, confidence refers to the

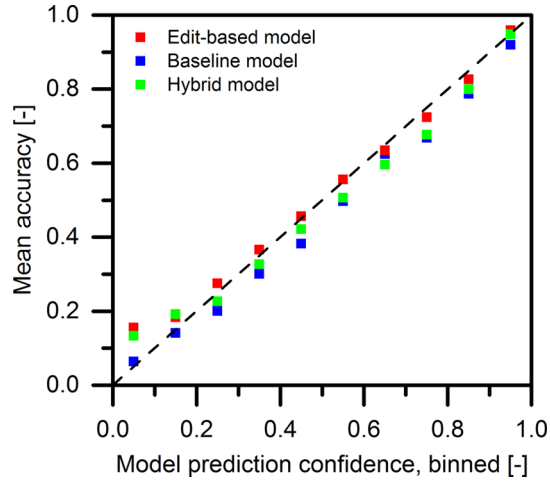


Figure 5. Mean model accuracy as a function of the binned model confidence, where the model confidence refers to the probability assigned to the highest-ranked candidate.

probability assigned to the highest-ranked candidate (i.e., the highest probability assigned to any candidate product). There is a very strong correlation of the model's accuracy with the prediction confidence; this signifies that the probability assigned to a candidate product does indicate the actual likelihood of that being the major product; this is a highly desirable characteristic for a predictive model. The distributions of prediction confidence are shown in Figure S10.

Prediction Examples. Analysis of individual predictions gives greater insight into model behavior than statistical measures. Figures 6 and 7 show the details of predictions from the hybrid model on test data.

Figure 6a depicts a functional group conversion from an alcohol to the corresponding chloride using thionyl chloride ($SOCl_2$). The recorded product is accurately predicted with an assigned probability of 94.8%. Sodium bicarbonate is also present, which introduces some competing reactivity channels (e.g., chlorination of bicarbonate), although these are all assigned a lower probability than the true outcome.

Figure 6b depicts a reaction between a carbamate and a secondary amine that leads to the carbamide, assigned a probability of 84.8%. The model recognizes that the tertiary amine is not a likely candidate for reactivity and that ethoxy is a plausible leaving group.

Figure 6c depicts a ring-forming reaction between an aryl alkyne and an iminol that leads to the isoxazole, assigned a high probability that rounds up to 100.0%.

Figure 6d depicts an S−N coupling between a primary arylamine and a sulfonyl chloride assigned a probability of 98.1%. Although sulfonamides can be prepared using secondary amines, the model recognizes that the diarylamine nitrogen is less reactive than the arylamine nitrogen. It also ranks the various possible Friedel−Crafts reactions (S−C coupling) lower.

Figure 6e depicts a simple S_N2 etherfication reaction between a phenol and an alkyl bromide, correctly predicted.

Figure 6f depicts a correctly predicted Suzuki Coupling between a pyridyl boronic acid and a pyridyl bromide. This outcome is assigned a probability of 98.8% and a rank of 1. Even though the necessary context (e.g., palladium catalyst) is missing and cannot be perceived by the model, the model

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443

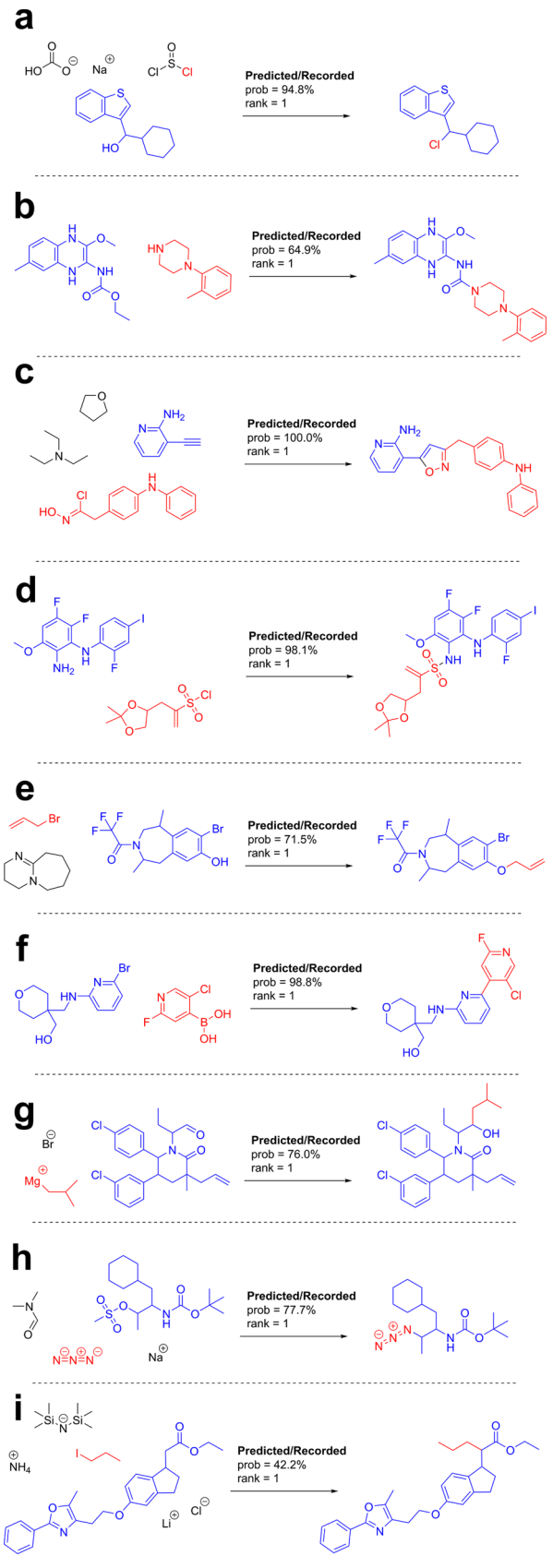


Figure 6. Reaction examples where the hybrid model assigned rank 1 to the recorded product. Recorded/predicted reactions: (a) chlorination; (b) amide synthesis; (c) isoxazole synthesis; (d) sulfamide synthesis; (e) etherification; (f) Suzuki coupling; (g) Grignard addition; (h) azidation; (i) alkylation.

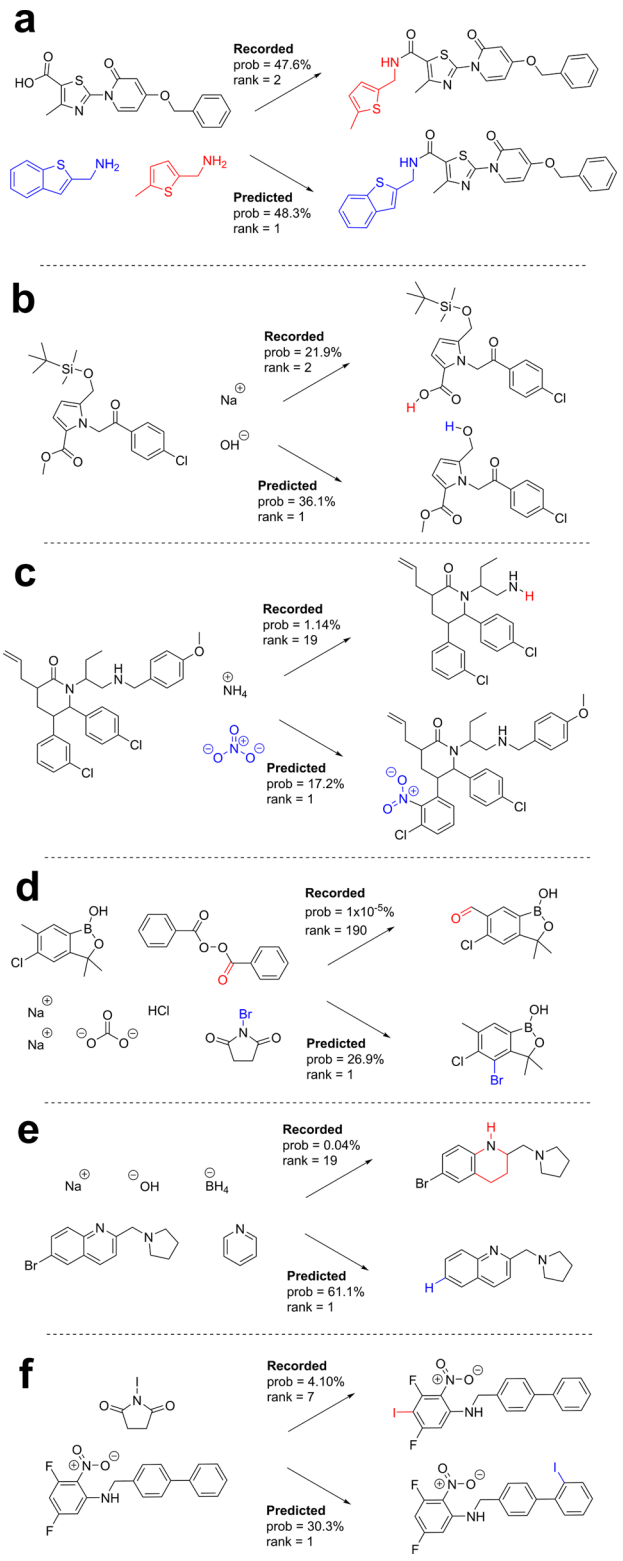


Figure 7. Reaction examples where the hybrid model did not assign rank 1 to the recorded product. Recorded [predicted] reactions: (a) amidation [amidation of different substrate]; (b) hydrolysis [hydrolysis at different ether]; (c) deprotection [nitration]; (d) oxidation [bromination]; hydrogenation [dehalogenation]; (f) iodination [iodination at different site].

implicitly assumes that the reaction would be be run under typical coupling conditions.

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443

Figure 6g depicts a Grignard reaction using isobutyl magnesium bromide. The alkylated substrate contains two potential carbonyl targets: an aldehyde and a nearby cyclic amide. The model correctly predicts the addition to occur at the aldehyde to form the secondary alcohol with high confidence.

Figure 6h depicts the preparation of an azide using sodium azide to replace the mesylate group. This recorded outcome is assigned a rank of 1 with a probability of 77.7%.

Figure 6i depicts a correctly predicted propylation. Propyl iodide reacts with the α carbon adjacent to a carboxylic ester. While the reaction may actually proceed through the enolate form, the model never explicitly constructs this intermediate species.

Turning to reaction examples for which the recorded product was not predicted correctly, Figure 7a depicts an amidation reaction between a carboxylic acid and a primary amine, where two separate primary amine substrates are available. It is likely that in the original document this example comes from, two separate reactions were reported for the two substrates, but these were somehow combined into one example during parsing and misrecorded in the database. Due to the similar properties at the reaction core (i.e., the two amines' nitrogen atoms), these two candidate outcomes are assigned very similar probabilities of 47.6% (recorded outcome) and 48.3% (alternate outcome); this produces a narrow misprediction. Remarkably, in the original CML file, the recorded product yield is 43%—quite close to our assigned probability. The uncertainty in our model is well-justified in this case.

Figure 7b depicts a reaction example where, in the recorded outcome, a methyl ester is hydrolyzed to form a carboxylic acid. On the same substrate, there is a *t*-butyldimethyl silyl ether, which is a commonly employed alkoxy protecting group highly resistant to hydrolysis. This resistance is not captured in our model, which mistakenly predicts the silyl ether to be the site of hydrolysis. The outcomes are assigned similar probabilities (23.6% and 36.1% with ranks 2 and 1), however, indicating that the model believes either of these two outcomes to be likely. This example highlights the importance of capturing sterics in candidate representations; in our atom-level featurization, two relevant attributes are included: the Labute Approximate Surface Area contribution and the Total Polar Surface Area contribution. Either the model has not seen enough examples in the training set where steric hindrance drives selectivity to understand its effects or—for this example—the perceived electronic differences between the two hydrolysis sites outweigh any perceived steric effects.

Figure 7c depicts a reaction where a secondary amine is dealkylated to form a primary amine in the presence of ammonium nitrate; this outcome is assigned a low probability of 1.14% and rank of 19, while the predicted outcome is assigned a still-low probability of 17.2%. In this case, because the ammonium counterion to nitrate is not present in the edit-based representation unless it is involved in the reaction, the model cannot distinguish weakly acidic conditions of the experiment from highly acidic nitric acid. The predicted outcome is thus an aromatic nitration on a substituted benzene ring ortho to chlorine.

Figure 7d depicts a reaction example where the recorded outcome, oxidation of a substituted toluene to form a benzaldehyde, is assigned an extremely low probability of 1×10^{-5}%. This reaction is plausible due to the strength of the oxidizing agent, benzoyl peroxide (BPO), but that is likely missed in our edit-based representation. The primary reason for this is that the radical mechanism by which BPO operates—and its corresponding reactivity—cannot be predicted based on the atom-level features used to describe the oxygen atom it contributes. Instead, the model predicts that *N*-bromosuccinimide (NBS) brominates the highly substituted aromatic ring. That prediction is consistent with the use of NBS as a brominating agent and is a plausible reaction outcome, particularly if BPO were not present. This example highlights the importance of order of reagent addition, which is not specified in the database used for training the model.

Figure 7e depicts a mispredicted reduction of a brominated quinolone species in the presence of sodium borohydride. In both the recorded and predicted outcomes, the presence of sodium borohydride is not captured as it does not contribute heavy atoms to the product molecule. The recorded hydrogenation of the pyridine ring is assigned a probability of 0.04% and a correspondingly low rank of 19. The model predicts instead the debromination with a relatively high probability of 61.1%, which has been reported for similar substrates.

Figure 7f depicts an iodination reaction by *N*-iodosuccinimide (NIS) where—due to the presence of three aromatic rings—there are many plausible reactive sites. The recorded outcome, assigned a rank of 7, is ortho to two fluorines, meta to a nitro group, and para to a secondary amine; these four substituents all direct reactivity to this site. While Gasteiger partial charges attempt to capture electron donating and electron withdrawing effects in aromatic rings, it is clear that the trained model does not properly capture the impact of directing groups. As with the previous example of silyl ether hydrolysis, this could be due to an insufficient input featurization or an insufficient number of reaction examples from which to learn this nuance.

Input Feature Analysis. To probe how the trained edit-based model depends on the input atom features, model performance is evaluated while masking certain feature indices (i.e., overriding the feature values of the candidate edits). Average values for each set of indices are calculated from the true candidates of the 10 500 reaction examples in the training set. These averaged values are then used during model evaluation on the test set. The resulting decrease in model performance measures how strongly the trained model relies on values of those attributes to discriminate candidates, although it does not reflect how the model might have developed had that attribute been missing during training. The results of this analysis for the edit-based model are reported in Table 2 in terms of relative test accuracy (averaged over the 5-fold CV). Masking tests were performed for each edit type individually.

The importance of atom-level attributes is not constant across the four edit types. Interestingly, the worst performance observed when masking a single input feature only represents a 5.2% decrease in accuracy. This signifies that the model is making use of information from many features and from all edit types to learn the nuances of chemical reactivity rather than relying on one or two attributes. For loss and gain of hydrogen, we see that the model relies on "functional" descriptors for evaluation, particularly the total polar surface area (TPSA) contribution, Estate index, Crippen contribution to molar refractivity (MR), and the Gasteiger partial charge. Loss or gain of a bond is also dependent on these features, but more so on the "structural" features describing the atomic number, number of neighbors, and number of hydrogens. Aromaticity is another important feature for assessing bond gain, likely due to the

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443

Table 2. Edit-Based Model Performance When Certain Input Features Are Set to Their Average Value for True Edits in the Training Set[a]

		type of edit masked, relative test accuracy			
ind.[b]	masked index (ces)	H loss (%)	H gain (%)	bond loss (%)	bond gain (%)
0	Crippen logP contribution	98.3	98.3	97.0	96.0
1	Crippen MR contribution	98.0	**96.6**	96.0	95.9
2	TPSA contribution	**96.6**	**96.4**	95.7	96.2
3	Labute ASA contribution	98.4	96.7	95.9	96.6
4	Estate index	97.8	**95.9**	95.6	95.9
5	Gasteiger partial charge	**97.1**	98.0	97.4	99.5
6	Gasteiger H partial charge	98.7	100.4	99.7	96.5
7−17	atomic number (1-hot)	98.7	98.2	**94.8**	96.8
18−23	number of neighbors (1-hot)	98.5	98.2	**94.9**	**95.4**
24−28	number of hydrogens (1-hot)	**97.8**	96.9	**95.2**	**95.6**
39	formal charge	100.4	100.2	100.2	100.4
30	is in ring	99.8	98.5	95.8	96.6
31	is aromatic	99.9	100.1	97.7	**95.4**

[a]Performance is reported in terms of relative accuracy on the test set. The three most significant features in each column are bolded for emphasis. [b]Indices in the $\mathbb{R}^{32}$ atom representation.

prevalence of aryl halide coupling reactions. These structural atom-level descriptors are similar to what have been used in convolutional molecular embedding.[24,25] There is significant room for improvement in identifying suitable atom-level descriptors that provide additional indications of reactivity or otherwise reflect the molecular context.

DISCUSSION

Forward Template Quality. An inherent shortcoming in the automatic extraction of reaction templates is their locality. Certain heuristic-driven techniques can be applied after template extraction[17] to improve applicability and consolidate many similar templates into fewer generalized ones, but this process is challenging without manual intervention. To begin to address the issue of locality, Soh et al.[26] describe an analysis of functional group occurrence in reaction examples to infer which groups promote/inhibit certain types of reactivity. While this is a more scalable approach than manually listing incompatible groups,[7] it has not been adapted into a directly translatable template-improvement algorithm.

Extracted templates are also highly reliant on atom-mapping to describe the correspondence between reactant and product atoms. Modern algorithms have evolved beyond simple maximum common substructure searches,[27,28] yet atom-mapping is certainly not a solved problem. In our approach, inaccurate atom-mapping is a less significant issue, as our templates are designed to be overgeneral to achieve high coverage of potential product species. The overall model performance does not depend strongly on atom mapping quality.

The coverage afforded by the forward templates affects the generalizability of the overall two-step model: for a yet-unseen reaction example, the model must be able to identify the true product as a candidate; otherwise the true product cannot be evaluated by the neural network. This same limitation is discussed by Segler and Waller.[12] Although they do not quantify the coverage of their automatically extracted 8720 templates, the average number of matches per query of 44.5 (compared to our 353) suggests correspondingly lower coverage. Morerover, our neural network model can evaluate *any* candidate reaction, even if the corresponding template and/or substrates have never been seen before. This makes the overall model highly extensible, as the template library can be expanded independently of neural network training. We do not observe a strong dependence of model performance on the number of candidates (Figure S11).

Candidate Reaction Representation. The vast majority of existing machine learning algorithms require fixed-length vectors or one-dimensional sequences as inputs. For single molecules, fixed-length vector representations include pharmacore fingerprints, Morgan or Extended-Connectivity fingerprints,[29] descriptor vectors, and—more recently—learned fingerprints.[24,25] Reactions could be represented as (a) the concatenation of reactant and product fingerprints, (b) the difference between reactant and product fingerprints, previously used for a reaction classification task,[30] or (c) some other representation with greater focus on the reaction core, like that of Kayala et al.[8,9]

In general, the performance of a neural network depends strongly on the choice of input representation. Representing reactions as concatenations of constituent molecules' fingerprints (optionally, reagents/catalysts too) proved viable in Wei et al.'s analysis of 16 reaction families for small compounds with 10 or fewer carbon atoms.[10] However, these types of molecule-level representation of reactants are insufficient for the types of real, multifunctional molecules used in organic synthesis; interactions between different combinations of substructures must all be considered. This is similar in philosophy to the combinatorial enumeration of possible mechanistic steps in Kayala et al.'s ReactionPredictor.[8,9]

The edit-based reaction representation is not without its flaws. From our analysis of which features the trained model is most dependent on, we find that the "functional" descriptors obtained through rapid molecule-level calculations are important in addition to the structural descriptors. With the current feature set, perhaps too much attention is placed on the reaction core, so important information about the nonreacting atoms is not sufficiently captured. There is opportunity for improvement in identifying suitable atom-level descriptors that provide additional indications of reactivity or otherwise reflect the molecular context.

Context Awareness. The current model assesses candidates only in terms of the molecules that contribute atoms to the final product, i.e., the reaction core. There is no consideration of reaction conditions, such as reactant concentration, catalyst identity, catalyst concentration, reagent concentrations, solvent(s), temperature, pressure, or reaction time. These factors can obviously have a tremendous effect on reaction outcome. In its current form, the lack of context means that the model weighs reaction candidates under implicitly defined typical conditions. The unavailability of contextual information also means that the model learns to limit the confidence with which it makes some predictions.

Preliminary unpublished work on context-aware models suggests that learning contextual effects will be challenging due to the sparseness of data. Even in commercial databases (e.g.,

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443

Reaxys) with partial manual curation, most reaction entries do not contain fully specified conditions. Moreover, reactions within a specific family are often reported under similar conditions, so extrapolation to atypical conditions would be challenging. Future work on context-aware models must address the question of how to train on and extrapolate from this limited data.

Reagent representation, a subproblem of context-awareness, poses another significant challenge. Representing a highly reactive reagent, e.g., a brominating agent, as a fingerprint or through a generic molecular representation is a poor use of information. From the reaction literature, we can identify the most popular agents for different additions (e.g., Br_2, NBS, PBr_3, CBr_4, LiBr for bromination), but it is not clear how to apply that knowledge in a scalable and fully automated way and capture that historical information in a vector representation. There are far too many reagents and catalysts to enumerate and encode each one in a one-hot manner, though this approach is feasible if confined to a small subset of chemistries.[10]

CONCLUSION

Using both a novel model framework for generating and ranking candidate reaction outcomes and a novel edit-based representation, we are able to reproduce in silico the qualitative results of actual experimental reactions. The unique framework combining candidate enumeration with ranking enables augmentation of existing reaction databases with hundreds of negative reaction examples implicit in every record. In a 5-fold CV of 15 000 such examples from the USPTO literature where the recorded product is found in a self-generated set of candidates, our hybrid model assigns the recorded product rank 1 in 71.8% of cases, rank 1−3 in 86.7% of cases, and rank 1−5 in 90.8% of cases; moreover, incorrect predictions are often chemically reasonable given the lack of detailed contextual information.

Through expanded atom- and bond-level featurization of the reaction core, especially descriptors with more direct relevance to chemical reactivity, this model framework can be expanded upon to achieve even higher predictive performance. And through the use of reaction examples with more complete information from commercial sources, a context-dependent model could be trained to understand reactivity in a more nuanced manner. There is a tremendous role for machine learning to play in computer assisted synthesis design, not only as a key component of automated retrosynthesis planning, but as a standalone tool for chemists to assess reaction viability. This work represents a significant step toward the long-term goal of virtual reaction screening and validation as a complement to experimental organic synthesis.

METHODS

All scripts were written in Python; RDKit[31] was used for molecule/reaction parsing, applying templates, and various cheminformatics calculations; Keras[32] using the Theano[33] backend was used for building the machine learning architecture.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscentsci.7b00064. Additional discussion, figures, and tables (PDF) All code used to produce the reported results can be found online at https://github.com/connorcoley/ochem_predict_nn. All data used are freely available and can be found via the same URL.

AUTHOR INFORMATION

Corresponding Authors

*(W.H.G.) E-mail: whgreen@mit.edu.
*(K.F.J.) E-mail: kfjensen@mit.edu.

ORCID

Connor W. Coley: 0000-0002-8271-8723
Klavs F. Jensen: 0000-0001-7192-580X

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the DARPA Make-It program under contract ARO W911NF-16-2-0023. C.W.C. received additional funding from the NSF Graduate Research Fellowship Program under Grant No. 1122374. The authors thank Joel Hawkins and Tim Jamison for commenting on the manuscript.

REFERENCES

(1) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178−192.
(2) Corey, E. J. General methods for the construction of complex molecules. *Pure Appl. Chem.* **1967**, *14*, 19−38.
(3) Pensak, D. A.; Corey, E. J. *Computer-Assisted Organic Synthesis*; ACS Symp. Ser.; 1977; Vol. *61*; Chapter 1, pp 1−32, doi:10.1021/bk-1977-0061.ch001.
(4) Salatin, T. D.; Jorgensen, W. L. Computer-assisted mechanistic evaluation of organic reactions. 1. Overview. *J. Org. Chem.* **1980**, *45*, 2043−2051.
(5) Satoh, H.; Funatsu, K. SOPHIA, a Knowledge Base-Guided Reaction Prediction System - Utilization of a Knowledge Base Derived from a Reaction Database. *J. Chem. Inf. Model.* **1995**, *35*, 34−44.
(6) Rose, P.; Gasteiger, J. In *Software Dev. Chem. 4, Proc. Workshop "Comput. Chem.", 4th*; Gasteiger, J., Ed.; Springer: Berlin, 1990; pp 275−288.
(7) Szymkuc, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904−5937.
(8) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to predict chemical reactions. *J. Chem. Inf. Model.* **2011**, *51*, 2209−2222.
(9) Kayala, M. A.; Baldi, P. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *J. Chem. Inf. Model.* **2012**, *52*, 2526−2540.
(10) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.
(11) Segler, M. H.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. - Eur. J.* **2017**, DOI: 10.1002/chem.201604556.
(12) Segler, M. H.; Waller, M. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, DOI: 10.1002/chem.201605499.
(13) ChemAxon, Reactor. https://www.chemaxon.com/products/reactor/, 2016.
(14) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *J. Chem. Inf. Model.* **2012**, *52*, 1745−1756.
(15) Lowe, D. M. Extraction of chemical structures and reactions from the literature. Thesis, University of Cambridge, 2012.

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443

(16) Lowe, D. M. Patent reaction extraction: downloads, 2014; https://bitbucket.org/dan2097/patent-reaction-extraction/downloads.
(17) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.* **2009**, *49*, 593−602.
(18) Boegevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Low, P.; Oppawsky, C.; Rein, T.; Saller, H. Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **2015**, *19*, 357−368.
(19) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464−477.
(20) Hall, L. H.; Mohney, B.; Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Model.* **1991**, *31*, 76−82.
(21) Gasteiger, J.; Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, *19*, 3181−3184.
(22) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity−a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(23) Zeiler, M. D. *ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701* 2012.
(24) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *NIPS* **2015**, 2224−2232.
(25) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.
(26) Soh, S.; Wei, Y.; Kowalczyk, B.; Gothard, C. M.; Baytekin, B.; Gothard, N.; Grzybowski, B. A. Estimating chemical reactivity and cross-influence from collective chemical knowledge. *Chem. Sci.* **2012**, *3*, 1497−1502.
(27) Fooshee, D.; Andronico, A.; Baldi, P. ReactionMap: An Efficient Atom-Mapping Algorithm for Chemical Reactions. *J. Chem. Inf. Model.* **2013**, *53*, 2812−2819.
(28) Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 560−593.
(29) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(30) Kraut, H.; Eiblmaier, J.; Grethe, G.; Low, P.; Matuszczyk, H.; Saller, H. Algorithm for Reaction Classification. *J. Chem. Inf. Model.* **2013**, *53*, 2884−2895.
(31) Landrum, G. RDKit: Open-source cheminformatics; http://rdkit.org. 2016.
(32) Chollet, F. keras; http://keras.io. 2015.
(33) Bastien, F.; Lamblin, P.; Pascanu, R.; Bergstra, J.; Goodfellow, I.; Bergeron, A.; Bouchard, N.; Warde-Farley, D.; Bengio, Y. *Theano: new features and speed improvements. arXiv preprint arXiv:1211.5590* 2012.

DOI: 10.1021/acscentsci.7b00064
ACS Cent. Sci. 2017, 3, 434−443